*Summer School on Information Extraction*, Frascati, July 14-19, Frascati, Italy.
To appear in a volume edited by M. T. Pazienza, published by Springer Verlag.

# Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration

Nicola Guarino
LADSEB-CNR, National Research Council
Corso Stati Uniti 4, I-35127 Padova
guarino@ladseb.pd.cnr.it
http://www.ladseb.pd.cnr.it/infor/Ontology/ontology.html

**Abstract**. The task of information extraction can be seen as a problem of semantic matching between a user-defined template and a piece of information written in natural language. To this purpose, the ontological assumptions of the template need to be suitably specified, and compared with the ontological implications of the text. So-called "ontologies", consisting of theories of various kinds expressing the meaning of shared vocabularies, begin to be used for this task. This paper addresses the theoretical issues related to the design and use of such ontologies for purposes of information retrieval and extraction. After a discussion on the nature of semantic matching within a model-theoretical framework, we introduce the subject of Formal Ontology, showing how the notions of parthood, integrity, identity, and dependence can be of help in understanding, organizing and formalizing fundamental ontological distinctions. We present then some basic principles for ontology design, and we illustrate a preliminary proposal for a top-level ontology develped according to such principles. As a concrete example of ontology-based information retrieval, we finally report an ongoing experience of use of a large linguistic ontology for the retrieval of object-oriented software components.

## 1. Introduction

With the dramatic increase of the possibilities of information access via the World-Wide-Web – paralleled by the constant increase of the costs of knowledge acquisition from scratch – the necessity of suitable tools for information organization, extraction, and integration has become more and more evident. In the "global information" perspective, the *added value* of a piece of coded information is no more only bounded to the particular application which motivated its acquisition, but tends to increase in dependence of its *reusability*, i.e. its suitability to be dynamically integrated within various, different bodies of information.

In this situation, the crucial characteristic of a piece of coded information is *what it is about*, i.e. the entities it refers to. It is this referential meaning that needs to be made

*Summer School on Information Extraction*, Frascati, July 14-19, Frascati, Italy.
To appear in a volume edited by M. T. Pazienza, published by Springer Verlag.

explicit and organized, in order the relevant information to be retrieved and used when necessary. It is easy to see how ontological aspects play a fundamental role here, and how the value of an information resource turns to be crucially bound to its *semantic transparency*. This is especially evident in the case of information extraction from text [Cowie and Lehnert 1996], where a user-defined template is to be "filled" by the information conveyed by a natural language statement. Here the problem is the *semantic matching* between the terms used to define the template and those appearing in the text: only if their meaning is clear, then it is possible to decide whether they match or not. In other words, the problem is to make explicit the *intended models* of the vocabulary used to convey and request information, clarifying the ontological assumptions implicit in the terms adopted for concepts, relations, attributes.

The purpose of this paper is to introduce the ontological notions underlying the problem of semantic matching, and discuss the emerging role of *ontologies*, i.e. theories of various kinds expressing the meaning of shared vocabularies, in the specific field of information retrieval and extraction as well as in the more general area of knowledge and language engineering. In the next section I analyze in some detail the nature of *semantic matching* within a model-theoretical framework, and the related notion of *conceptualization*. After a clarification of the meaning of the term "ontology" in the current practice of knowledge engineering (Section 3), I introduce in Section 4 the subject of formal ontology, showing how the basic notions of parthood, integrity, identity, and dependence can be of help in understanding, organizing and formalizing fundamental ontological distinctions. In Section 5 I propose some basic principles for ontology design, while in Sections 7 and 8 I present a preliminary proposal for the top-level ontology of, respectly, *particulars* (entities of the world) and *universals* (relations involving entities of the world). Finally, as a concrete example of ontology-based information retrieval, I report in Section 9 an ongoing experience of use of a large linguistic ontology like SENSUS [Knight and Luk 1994, Swartout *et al.* 1996] for a project of ontology-based retrieval of object-oriented software components.

Most of the material presented here is an adaptation and a systematization, with didactic purposes in mind, of work already appeared in previous papers. References to the relevant original works appear throughout the paper. Sections 6 and 7, presenting a concrete proposal of a top-level ontology, are rather sketchy and compact, partly for the sake of conciseness, but more importantly because they refer to very recent work, previously introduced in a preliminary form in [Guarino 1997a].

## 2. Semantic Matching

Just to keep things simple, suppose you want to produce a list of natural language descriptions satisfying a certain property, say Red(X), by automatically extracting them from a body of text. You use a lexical item "red" to describe your template, but, without making clear what its *intended meaning* is, the system can only perform a syntactic matching, without distinguishing between its different meanings (Fig. 1).

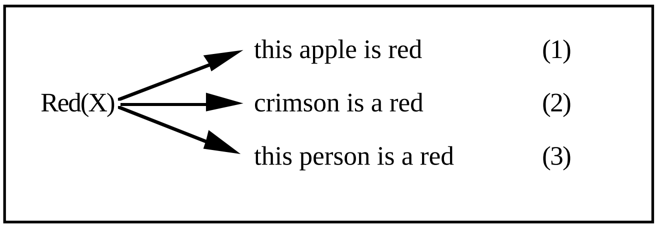


**Fig. 1**. If the *intended meaning* of the template is not specified, problems of semantic match arise. (From [Guarino *et al.* 1994]).

Let us try to state the problem in more general terms. Consider a logical language **L** using a certain set **V** of predicate symbols, called the *vocabulary* (or the *signature)* of that language. When an agent **A** uses **L** for some purpose, the *intended models* of **L** according to **A** will constitute a small subset of the set **M**(**L**) of all models of **L** (Fig. 2). We call such set of intended models the *conceptualization* of **V** according to **A**.

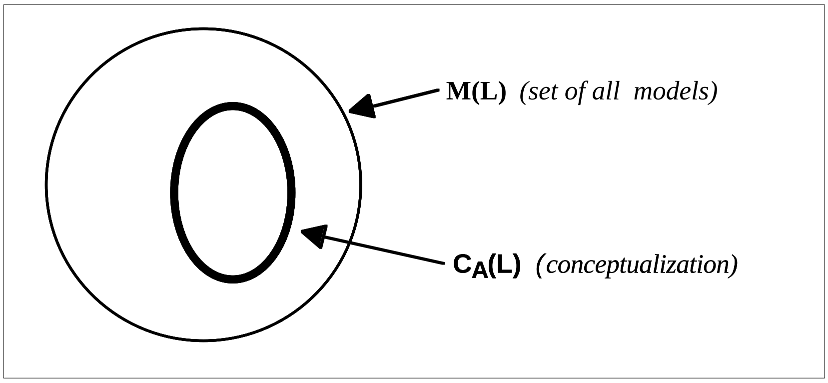


**Fig. 2**. $\mathbf{C}_A(\mathbf{L})$ is the *conceptualization* of the language **L** according to the agent **A**. It is a subset of the set **M**(**L**) of all models of **L**, representing those models (the *intended models*) which are compatible with the intended meaning of the vocabulary used.

Consider now two different agents, **A** and **B**, using the same language **L**: in order to give the same meaning to the vocabulary used they must either already share the same conceptualization, or otherwise agree on adopting a common conceptualization which is the intersection of the two distinct original conceptualizations (Fig. 3). In the context of information extraction, the process of semantic matching between the template and the data implies this kind of agreement. As we shall see, the key role of ontologies in information extraction is to help establishing this agreement.

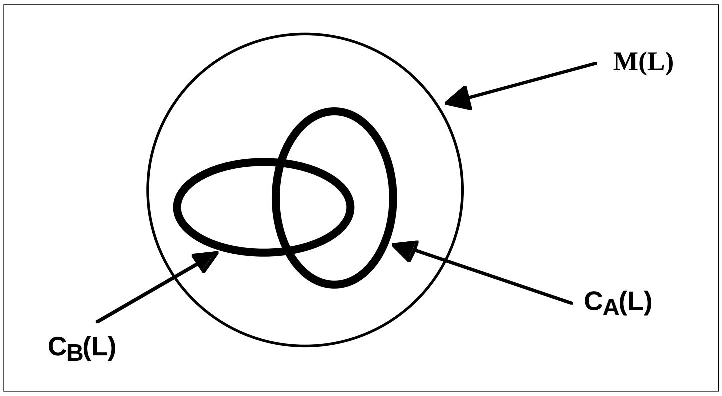


**Fig. 3.** Two agents **A** and **B** using the same language **L** can communicate only if their conceptualizations $\mathbf{C}_A(\mathbf{L})$ and $\mathbf{C}_B(\mathbf{L})$ overlap.

## 3. What is an Ontology

But what is exactly *an* ontology? How does it differ from Ontology, with the capital O, intended as a branch of philosophy? Together with Pierdaniele Giaretta, I have addressed this question in [Guarino and Giaretta 1995]; I report here some of the discussion appeared there (also further developed in [Guarino 1997b]).

In the AI literature, an ontology (intended as an engineering artifact, constituted by a particular piece of code) has been defined as a "specification of a conceptualization" [Gruber 1995]. In the model-theoretical framework introduced above, this definition can be interpreted according to Fig. 4 below.

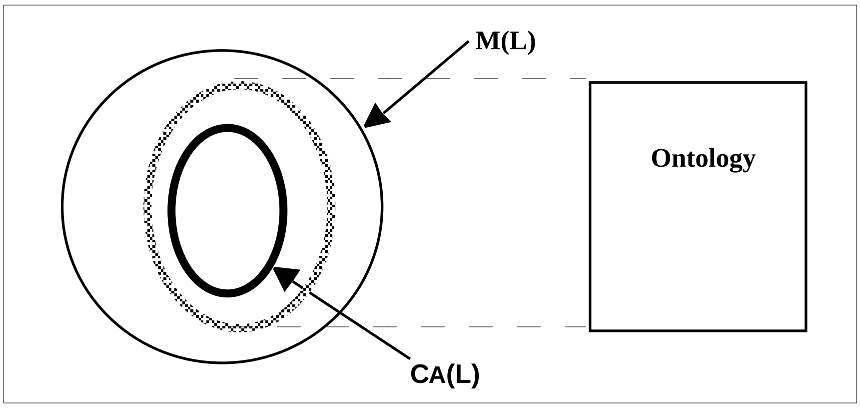


**Fig. 4**. An ontology characterizes a certain conceptualization of a language by constraining the intended models of such language

An ontology, according to such an interpretation, is a logical theory whose models *constrain* a particular conceptualization, whithout exactly *specifying* it. In other words, an ontology is an axiomatic characterization of the *meaning* of a logical vocabulary. In many cases, the axioms of an ontology only express subsumption (ISA) relationships between unary predicates, but of course a more detailed axiomatization is often necessary in order to exclude unwanted interpretations.

### 3.1 What is a conceptualization.

The notion of conceptualization I have introduced here requires some further comments, since some confusions about this term exist in the literature. In particular, Gruber's definition of ontology reported above explicitly refers to the notion of "conceptualization" introduced by Genesereth and Nilsson in their well-known text-book on AI [Genesereth and Nilsson 1987]; such a notion is however more akin to that of "state of affairs", while Gruber seems to appeal to the intuitive meaning of the term "conceptualization". Let us consider the example given by Genesereth and Nilsson. They take into account a situation where two piles of blocks are resting on a table (Fig. 5a). According to them, a possible conceptualization of this scene is given by the following structure:

$$<\{a, b, c, d, e\}, \{on, above, clear, table\}>$$

where $\{a, b, c, d, e\}$ is the *universe of discourse*, consisting of the five blocks we are interested in, and $\{on, above, clear, table\}$ is the set of the relevant relations among this blocks, of which the first two, *on* and *above*, are binary and the other two, *clear* and *table*, are unary. The authors make clear that objects and relations are extensional entities. For instance, the *table* relation, which is understood as holding of a block if and only if that block is resting on the table, is just equal to the set $\{c, e\}$. It is exactly such an extensional interpretation that originates our troubles.

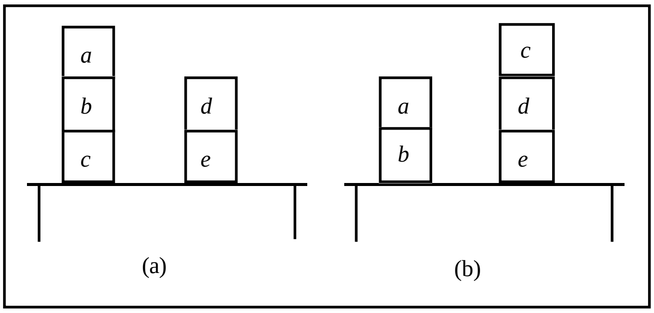


**Fig. 5**. Blocks on a table. (a) A possible arrangement of blocks. (b) A different arrangement. Also a different conceptualization? (from [Guarino and Giaretta 1995])

Referring to the example given, consider a different arrangement of blocks, where *c* is on the top of *d* and *a* and *b* form a separate stack standing on the table (Fig. 5b). The corresponding structure would be different from the previous one, generating therefore – according to Genesereth and Nilsson – a different "conceptualization". Of course there is nothing wrong in such a view, if one is only interested in isolated snapshots of the world. But the meanings of the terms used to denote the relevant relations are still the same, since they are invariant with respect to the possible configurations of blocks. In fact, in the metalanguage adopted in their book, Genesereth and Nilsson would adopt the same symbols (*on*, *above*, *clear*, *table*) to denote the new conceptualization. We prefer to say in this case that the *states of affairs* are different, but the conceptualization is the same, as it is related to *the way* we assign a particular relation to each symbol in the new situation.

Formally, the meaning of a symbol like *on*, for instance, can be seen as a function $f: W \rightarrow 2^{D^2}$, where W is the set of all possible states of affairs and D={*a*, *b*, *c*, *d*, *e*} is the universe of discourse. In other words, a conceptualization expresses the intended meaning of each symbol independently of the particular situation at hand: the rules which tell us whether a certain block is *on* another one remain the same, independently of the particular arrangement of the blocks. These rules can be expressed by suitable axioms, like for instance $\neg on(x,x)$, which constrain the intended models in the sense described in Fig. 4.

### 3.2 Ontologies and knowledge bases

Having made clear that an ontology corresponds to a logical theory, the obvious question arising is the difference between an ontology and a knowledge base. The answer is related to the *purpose* of an ontology, which is a *particular* knowledge base, describing facts assumed to be *always true* by a community of users, in virtue of the agreed-upon meaning of the vocabulary used. A *generic* knowledge base, instead, may also describe facts and assertions related to a particular state of affairs or a particular epistemic state.

We can classify ontologies according to two dimensions: their *level of detail* and their *level of dependence* on a particular task or point of view. On the first dimension, a very detailed ontology gets closer to specifying the intended meaning of a vocabulary (and therefore may be used to *establish consensus* about sharing that vocabulary, or a knowledge base which uses that vocabulary), but it usually pays the price of a richer representation language. A very simple ontology, on the other hand, may be developed with particular inference services in mind, in order to be shared among users which *already agree* on the underlying conceptualization. We can distinguish therefore between *reference ontologies* and *shareable ontologies*, or maybe *off-line* and *on-line ontologies*. Very simple ontologies like thesauri can be kept on-line, while sophisticated theories accounting for the meaning of the terms used in a thesaurus can be kept off-line.

On the second dimension, we may distinguish between *top-level ontologies, domain ontologies, task ontologies* and *application ontologies*: (Fig. 6)

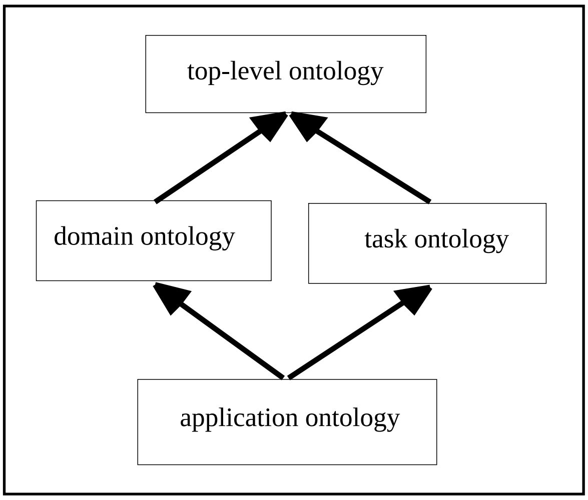


**Fig. 6**. Kinds of ontologies, according to their level of dependence on a particular task or point of view. Thick arrows represent specialization relationships.

- *Top-level ontologies* describe[1] very general concepts like space, time, matter, object, event, action, etc., which are independent of a particular problem or domain: it seems therefore reasonable, at least in theory, to have unified top-level ontologies for large communities of users.
- *Domain ontologies* and *task ontologies* describe, respectively, the vocabulary related to a generic domain (like medicine, or automobiles) or a generic task or activity (like diagnosing or selling), by specializing the terms introduced in the top-level ontology.
- Application ontologies describe concepts depending both on a particular domain and task, which are often specializations of both the related ontologies. These concepts often correspond to *roles* played by domain entities while performing a certain activity, like *replaceable unit* or *spare component*.

The interested reader may refer to [Uschold and Gruninger 1996, Van Heijst *et al.* 1997] for a general introduction on the use of ontologies in the practice of knowledge engineering, and [Guarino 1997b, Van Heijst *et al.* 1997] for an account of the current debate about the role and the nature of task ontologies and application ontologies.

### 3.3 Ontologies and KR languages

A further, separate kind of ontology is constituted by what have been called *representation ontologies* [Van Heijst *et al.* 1997]. They are in fact meta-level ontologies, describ-

[1] I prefer to use the verb “describe” rather than “define”, since very rarely it is possible to completely *define* the meaning of a term.

ing a classification of the primitives used by a knowledge representation language (like concepts, attributes, relations...)[2]. An example of a representation ontology is the *Frame Ontology* [Gruber 1993], used to support translations within different knowledge representation languages. A further example is the ontology of meta-level primitives presented in [Guarino *et al.* 1994], which differs from the Frame Ontology in assuming a non-neutral ontological commitment for the representation primitives. Such a position has been further discussed in [Guarino 1994, Guarino 1995], where the notion of *ontological level* for knowledge representation languages has been introduced.

## 4. The Tools of Formal Ontology

In the current practice, the process of ontology building lacks well-established principles and methodologies. In the past, ontological issues have been rather neglected in computer science and especially in artificial intelligence, where reasoning and problem-solving have been largely emphasized with respect to knowledge representation and conceptual analysis. Some methodological proposals have been made in recent times [Bateman *et al.* 1990, Bouaud *et al.* 1995, Gruber 1995, Mahesh 1996, Uschold and Gruninger 1996], but most of the ontologies currently in use seem to be the result of a mixture of ad-hoc creativity and naive introspection. Recently, however, a different line of research began to emerge, characterized by a highly interdisciplinary perspective: while staying on the solid grounds of logic and computer science, it shows an open-minded aptitude towards the subtle distinctions of philosophy and the slippery issues of natural language and commonsense. The philosophical field inspiring this trend is that of *formal ontology*, which has been defined as “the systematic, formal, axiomatic development of the logic of all forms and modes of being” [Cocchiarella 1991]. As such, formal ontology is a recent expression of traditional ontology, intended as the branch of philosophy whih deals with the *a priori* nature of reality[3]. In its current shape, formal ontology can be seen as the confluence between a school of thought which has addressed metaphysical problems within the mainstream of analytic philosophy, and another school more closely related to phenomenology, in the tradition of Brentano and Husserl. The former school includes a multitude of philosophers, which roughly agree on the idea of "descriptive metaphysics" proposed by Strawson [Strawson 1959, Aune 1991]. The latter sees the philosophers of the so-called "school of Manchester" [Smith 1982, Smith and Mulligan 1983, Simons 1987, Mulligan 1992] as its principal defenders. The reader

---

[2] Representation ontologies are therefore special cases of ontologies of universals (Section 7).

[3] The genuine interpretation of the term “formal ontology” is still a matter of debate (see for instance [Poli 1995]): in the Husserlian sense,“formal” has to be intended as opposite to “material”, and means therefore “independent of a particular content”; here the adjective is intended as synonymous of both “rigorous” and “a priori”, including both “analytic a priori” and “synthetic a priori”. For instance, the properties of matter and space would be considered as “material” in the Husserlian sense, while they are included in the scope of the notion of formal ontology adopted here.

may refer to [Guarino and Poli 1995] for an overview of the role of formal ontology in the information technology. For a general reference on metaphysics and ontology, see [Burkhardt and Smith 1991].

In practice, formal ontology can be seen as the *theory of a priori distinctions* within:

- (our perception of) the entities of the real world, or *particulars* (physical objects, events, regions of space, amounts of matter…)
- the categories we use to talk about the real world, or *universals* (concepts, properties, qualities, states, relations, roles, parts…)

The study of formal ontological distinctions can be organized around a number of core theories (or, better, theoretical tools), which have always been at the basis of philosophical research. I describe here briefly some of these theories, which in my opinion may contribute to establish a solid and fruitful ground for ontological analysis.

**Theory of parts**. A theory of parthood is at the basis of any form of ontological analysis. Relevant questions that must be addressed are:

- What counts as a part of a given entity?
- What are the properties of the parthood relation?
- Are there different kinds of parts?

An important example of a theory of parthood is given by *extensional mereology*. Much work must be addressed however in order to come up to a satisfactory theory of *intensional mereology*, where integrity and identity are taken into account. See [Simons 1987] for a thorough reference to the problems of mereology.

**Theory of wholes** (or theory of *integrity*). A given entity (like a collection of disconnected pieces) can have parts without being considered as a single whole. The theory of wholes studies the ways of connecting together different parts to form a whole. Relevant questions that must be addressed are:

- What counts as a whole? What does make it a whole?
- In which sense are the parts of whole *connected*? What are the properties of such connection relation?
- How is the whole isolated from the background? What are its boundaries?
- What is the role played by the parts with respect to the whole?

Notice that in order to understand the various forms of part-*whole* relation [Winston *et al.* 1987, Artale *et al.* 1996] the general theory of parthood must be supplemented with a theory of integrity. Together, the two theories form what may be called *mereotopology* [Varzi 1996].

**Theory of identity**. The theory of identity builds up on the theory of parthood and the theory of wholes, studying the conditions under which two entities exhibiting different properties can be considered as the same. Relevant questions that must be addressed are:

- How can an entity change while keeping its identity?
- What are its essential properties?
- Under what conditions does an entity loose its identity?
- Does a change of parts affect identity?
- Does a change of "point of view" change the identity conditions?

The last question is especially relevant in our case to distinguish between ontological strata (see below). For instance, consider the classical example of a vase of clay. Should we consider the vase and the clay it is made of as two separate individuals, or just as two different points of view about the same individual? The answer may be difficult, but a careful analysis tells us that the two views imply different identity criteria: when the vase looses its identity by crashing to the floor, the clay is still there. This is because the clay has an *extensional* criterion of identity, since it always coincides with the sum of its parts, while the vase requires a particular arrangement of its parts in order to be a vase. Therefore, we are in presence of two different individuals. A rigorous analysis of this argument requires some care, of course, but the example gives the idea. See [Hirsch 1982] for an account of the identity problems of ordinary objects, and [Noonan 1993] for a review of the philosophical research in this area.

**Theory of dependence**. The theory of dependence studies the various forms of existential dependence involving specific individuals belonging to different classes. We refer here to the notion of existence as "concrete existence", not as "logical existence". In this sense, existence can be represented by a specific predicate (like in [Hirst 1991]) rather than by a logical quantifier. Relevant questions that must be addressed are:

- Does the concrete existence of an individual necessarily imply the concrete existence of another specific individual? *(Rigid dependence)*
- Does the concrete existence of an individual necessarily imply the concrete existence of some individual belonging to a specific class? *(Generic dependence)*
- Does the fact that an individual belongs to a particular class necessarily imply the fact that another individual belongs to another class? *(Class dependence)*

An example of rigid dependence may be the relationship between a person and his/her brain, while the relationship between a person and his/her heart is an example of generic dependence (because the heart can be substituted with another heart, and the identity of the person does not change). Finally, an example of class dependence is the relationship existing between the class "Father" and the class "Child". In our opinion, the notion of dependence is at the core of Peirce's distinction between Firstness, Secondness and Thirdness, proposed in [Sowa 1995]. However, the examples reported by Sowa are far from being clear, mainly because they don't take into account the distinction between particulars and universals (see below).

**Theory of universals**. According to Aristotle, *particulars* are entities that “cannot be said of anything” (except themselves); they correspond to individuals existing either in the actual or in a possible world. *Universals*, on the other hand, are entities which “can be said of something”, usually corresponding to classes and relations. If these classes and relations are *reified*, and taken therefore as elements of a domain, the theory of universals can be seen as a sort of meta-level ontology, which builds on the theory of particulars to introduce useful distinctions among universals.

Notice however that what we usually call class and relations cannot have – in an ontological setting – the standard semantics of sets of tuples, since their *meaning* is supposed to be independent on particular states of affairs (worlds). For these reasons, we give them a Montague-style intensional semantics, as discussed in [Guarino and Giaretta 1995]. A relation of arity n is not just a set of tuples, but rather a function from a set of possible worlds to the set $2^{D^n}$, where D is the domain where the relation is defined.

The following formal properties of classes turn to be of particular relevance for ontological purposes[4]:

- *Countability*: A class is countable if, for all of its instances, none of their proper parts is an instance of the same class [Griffin 1977] (“Person” is countable, “Water” is not)
- *Rigidity*: A class C is rigid if, when C*x* is true, then it must be *always* true (“Person” is rigid, “Student” is not).

## 5. Some Basic Principles for Ontology Design

After this review of the core theories of Formal Ontology, let us show now how these theories can be exploited in order to establish a methodological foundation for ontology design, and to enlighten the fundamental distinctions that must be recognized in every task involving the analysis and the organization of a body of information. In the field of information retrieval and extraction, the study of these distinctions may help developing less user-dependent ways of organizing information and making explicit the intended meaning of linguistic terms.

A common problem of many current top-level ontologies, like CYC [Lenat and Guha 1990], PENMAN/PANGLOSS [Bateman *et al.* 1990, Knight and Luk 1994] or Mikrokosmos [Mahesh 1996], is that they present a tangled taxonomic structure, so that it is difficult to isolate a basic backbone, a sort of natural skeleton or “conceptual coat rack” [Woods 1986] to be used for cognitive and organizational purposes. In the follow-

[4] The definition reported are simplified. See [Guarino *et al.* 1994] for a more detailed account.

ing, I present a set of principles I have used to isolate a preliminary taxonomy of top-level ontological concepts aimed to conciliate clarity with semantic rigour, generality and commonsense, which *results* to be a tree, although being a tree is not its most important (nor especially desired) property. These principles can be summed up as follows:

1. Two distinct ontologies are assumed for *particulars* and *universals*.
2. Not all the unary relations appearing in the ontology of universals correspond to *explicit* concepts in the top-level ontology of particulars. Such concepts are called *taxons*, and must be chosen according to suitable organization criteria.
3. Taxons correspond to what are called in [Strawson 1959] *sortal* categories, i.e. unary relations which imply a specific identity criterion for their instances (like *physical body* and *event*). Examples of *non-sortal* categories are *decomposable entity* or *red thing* , which do not carry a specific identity criterion.
4. Taxons are all primitive. Formal ontological properties (many of whom are defined) contribute to characterize their meaning, in terms of necessary conditions.
5. Taxons are mainly organized according to their specific identity criteria. Different identity criteria correspond to disjoint taxons. Identity criteria can be grouped in classes, corresponding to a stack of *ontological strata* linked together by a dependence relationship [Borgo *et al.* 1996].

In summary, I argue that a top-level ontology should not explicitly represent as concepts all the relevant properties, but only those corresponding to *sortal* properties; the *non-sortal* properties, which contribute however to characterize the meaning of concepts, can be represented by means of slots and axioms. The result of this choice, together with the application of the principle of stratification according to different kinds of identity criteria, is such that the structure of the top-level becomes considerably simplified, guaranteeing at the same time a rigorous semantics.

## 6. The Ontology of Particulars

In the following, I shall first introduce what I consider as a basic "backbone" of distinctions among particulars, organized around sortal categories; then I present three possible dimensions of non-sortal categories which can be used to characterize and further specialize the basic backbone. The first of these dimensions introduces a stratification on the basic backbone on the basis of identity criteria, while the other two are based on considerations of integrity and dependence. A preliminar, schematic top-level for the ontology of particulars is reported in Appendix 1.

### 6.1 Locations, Substrates, Objects, and Qualities

The basic "backbone" of the distinctions I propose is reported in Fig. 7. All these categories are assumed as primitive sortal categories. I give in the following an informal account of their properties, and the relationships among them. A preliminary formal account, limited to the case of space, matter, and physical continuants, appears in

[Borgo and Guarino 1997]. Just to visualize the perspective I have in mind, suppose to have a workshop for the production of plasticine artifacts: you have a lump of plasticine on one side, a clean table where to put your artifacts, a side desk with colors and a clock in front of you. We are interested in modelling a "world" which is the table area, where various objects may appear or disappear at different times: for instance, we may decide to destroy an object in order to reuse its plasticine later, or to substitute some of its pieces.

| | |
|---|---|
| Particular | |
| Substrate | |
| Location | |
| Space | *(a region of the table)* |
| Time | *(a time clock)* |
| Matter | *(an amount of plasticine)* |
| Object | |
| Concrete object | |
| Continuant | *(a plasticine artifact)* |
| Occurrent | *(the movement of an artifact on the table)* |
| Abstract object | *(Pitagora's theorem)* |
| Quality | *(the color of a particular piece of plasticine)* |

**Fig. 7**. The basic "backbone" in the ontology of particulars.

Notice that we distinguish the *actual existence* of an object on the table from its *logical existence* as a possible artifact. Its actual existence is a matter of contingency, while its logical existence is constrained by the "laws of nature". This assumption is needed in order to classify objects according to their behavior across possible worlds, avoding at the same time to adopt a modal logical framework. The aim is not only that of keeping things simple from the technical point of view, but also that of offering a graspable notion of the "possibility" dimension, which appears to be unavoidable in any ontological endeavour.

*Substrates* are the ultimate entities on which objects depend. Their main characteristic is that they have an *extensional identity criterion*. Matter is the classical example of a substrate in the Aristotelian sense. I include here also *locations* like space and time within this category, since they present similar characteristics for what concerns their identity criteria and their depedence relationships[5].

A *location* is either a region of (absolute) space or an interval of (absolute) time. I do not enter here in the debate between regions and points: we can either assume regions are sets of points or adopt a "pointless" topology thinking of "points" as regular regions of minimal granularity. See [Borgo *et al.* 1996] for a formal account of the latter approach

---

[5] I don't exclude other possible substrates; for instance , it is not clear to me whether abstract objects have a substrate, too (knowledge?).

in the domain of space.

Regarding *matter*, notice that I do not assume that an amount of matter has always a spatial location: it acquires *actual existence* when it gets a spatial location in a specific interval of time (i.e. when, in our example, a certain amount of plasticine is taken out from the lump and put on the table). In this way we can quantify over different worlds having a varying total amount of matter.

*Concrete objects* (or simply "objects", if the context is clear) are intended here in a very general sense, to refer to entities which can be located in space and/or in time and somehow depend on a certain amount of matter, like me and my typing on the computer right now. They are characterized by the way they depend on other objects, and the way they *can* extend in space and time. As in the case of matter, they acquire actual existence (distinguished from logical existence) when they get a location in spacetime.

Within concrete objects, I assume here for granted the classical distinction between *continuants* and *occurrents*. These two terms are less ambiguous than other terms commonly used such "object", "event", "process", or "eventuality".

In order to have actual existence, continuants must get a location in space, but this location can vary with time. They can have spatial parts, but they do not have a temporal location, nor temporal parts: my hand is a part of my body, but my youth is not part of me: it's part of my life. I am not part of my life, but rather I *take part* to my life... In other words, I refuse to admit that continuants have a *location* in time coinciding with their temporal existence, since I want to say that they *fully exist* also in each time interval within their life. All continuants *depend* on matter (which is considered therefore as their *substrate*), since their identity criteria are bound to the ways matter exhibits specific properties (in particular, mereo-topo-morphological properties).

Occurrents are "generated" by continuants, according to the ways they behave in time. Examples of occurrents are the change of location of a piece of matter (an *event* which extends on a given time interval), but also the permanence of a piece of matter in a given location for a given time (a state occurrence, i.e. a *static event*). Occurrents depend on continuants, but in a way different from the way continuants depend on their substrate, since we can say that matter *constitutes* physical bodies, while the latter *take part* to occurrents whithout constituting them nor being parts of them. Occurrents have temporal parts (they *perdure* in time), while continuants do not (they *endure* in time). Occurrents always have a unique temporal location, while their exact spatial location is less obvious, and it is however bound to the location of the participating continuants. In the following, when mentioning the location of an occurrent, I shall refer to its temporal location. For a thorough review of the philosophical work on occurrents, see [Casati and Varzi 1996].

*Abstract objects* are included here just for completeness, but are in fact kept outside the current analysis. Notice however that I refer here to abstract *particulars*, such as Pitagora's theorem or (maybe?) the number "1". Most entities usually called abstract objects are however *universals* (see below).

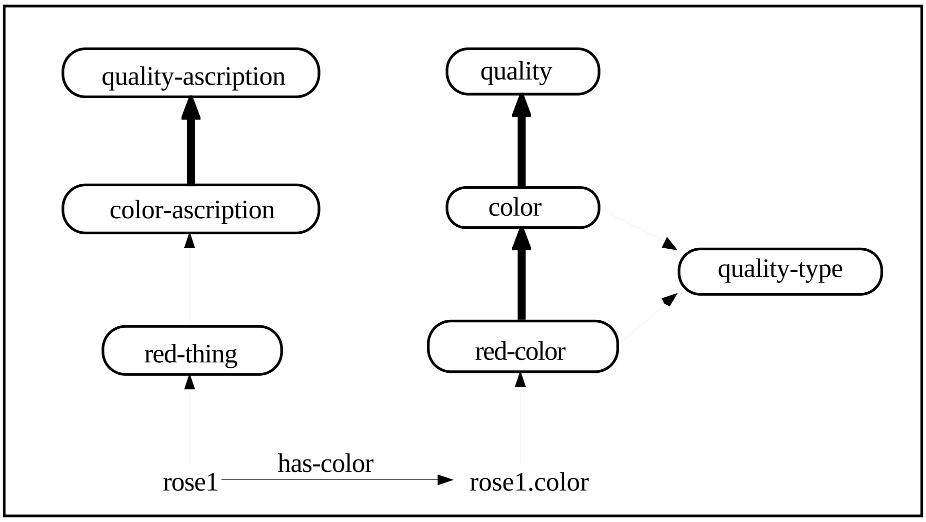


**Fig. 8**. *rose1.color* is a quality of *rose1*. Both are considered as particulars. *red-color* is a class of qualities having certain common characteristics, while *red-thing* is the class of particulars exhibiting such characteristics. Such a class is an instance of the metaclass *quality-ascription*, while *color* and *red-color* are both instances of the metaclass *quality-type*. Thin unlabelled arcs denote instance-of relationships, while thick arcs subsumption relationships.

*Qualities* are included here among particulars. I endorse here the position of those philosophers who – following Husserl – admit the ontological existence of "individual qualities" [Mulligan *et al.* 1984]. Therefore the color of a particular rose, the height of a particular mountain, the length of my nose are all unique individuals, not just properties like "being red" or "being 2.5 cm. long". If $x$ is a rose, it is *its* color, not the color of another rose, which is responsible of making the proposition Red($x$) true. Such a color is an individual dependent on the rose itself, but different from it. Briefly, the opportunity of admitting individual qualities comes from:

- the presence of distinct perception stimuli;
- the difficulties of giving a formal account to sentences like "the color of this rose turned from red to purple in two hours";
- the naturality of the individual interpretation for expressions like "the area of New York", "the volume of a body", "the velocity of a body" (intended as an applied vector);
- the difficulties of representing the relationship between specific properties like having a particular speed and "property spaces" like "speed" if this is intended as the set of all such properties, especially in presence of multiple scales of measurement.

Consider for example the case of a particular rose, *rose1* (Fig. 8). Its color (*rose1.color*) can be modeled as an instance of *red-color*, which is a subclass of *color*, which itself is subclass of *quality*. *Rose1* will be then an instance of *red-thing*, which is an instance of the metaclass *color-ascription*, which itself is a subclass of *quality-ascription*. We can explain in this way the relationship between *color*, *red-color* and *red-thing*. Recalling the example of Fig. 1, the meaning of "red" in statement 1 is that of "red-thing", while the meaning of the same term in statement 2 is that of "red-color". Instances of quality types correspond to what have been called *determinable* properties, while instances of quality ascriptions correspond to *determinate* properties [Cleland 1991]. Consider another example of determinable, like "speed". It turns to be just a class of individual qualities, while each set of determinates relative to a particular scale, like {"1 m/s", "2 m/s", ...} forms a *gradient* of mutually exclusive subclasses of "speed". These gradients have been called "quality spaces" in [Lombard 1979]. Under this approach, a *unit of measure* is nothing more that a specific individual quality taken as a reference. A relevant relation between qualities is the *congruence* relation, which allows us to *measure* an individual quality by comparing it with a reference quality.

### 6.2 Ontological Strata

Let us now introduce a set of *non-sortal* categories useful to establish further distinctions and characterizations in the main backbone described above, on the basis of considerations bound to the theory of identity and the theory of dependence. The idea of *strata* of reality has been inspired by [Poli 1996, Albertazzi 1989]. The notion adopted here, introduced in [Borgo *et al*. 1996], is however more explicitly related to the theory of identity.

I have already underlined that *substrates* have an extensional criterion of identity, in the sense that any two of them are identical if they have the same parts; objects, on the other hand, have an *intensional* criterion of identity, in the sense that they are more that mere sums of parts. Within objects, further distinctions can be made according to the identity criteria ascribed to them. For instance, an animal can be conceptualized as an intentional agent, as a biological organism or just as a piece of matter. I argue that different identity criteria correspond to *disjoint* concepts: in the case of an animal, three distinct individuals, corresponding to the three conceptualizations above, share a common spatial location. Since the animal depends on the underlying biological organism, as well as the biological organism depends on the underlying amount of matter, I call these different sub-concepts *ontological strata.*

The proposed classification of ontological strata is reported below (Fig. 9). Strata correspond to different *kinds* of identity criteria (IC) for particulars, distinguished according to the properties used to express them. A *dependence relation* links higher strata to lower strata: an animal depends on its body, which depends on body parts having a certain functionality, which depend on pieces of matter having specific properties, and so on. Notice that this dependence is of a *generic* kind: a vase depends on *some* amount of clay, but not necessarily always the same specific clay (it can loose parts, it can be repaired with a different piece of clay).

| | |
|---|---|
| Ontological stratum | |
| Mereological | *(an amount of matter)* |
| Physical | |
| Static | *(a state or a state occurrence)* |
| Topological | *(a piece of matter)* |
| Morphological | *(a cubic block)* |
| Functional | *(an artifact)* |
| Biological | *(a human body)* |
| Intentional | *(a person or a robot)* |
| Social | *(a company)* |

**Fig. 9**. Ontological strata describe *disjoint sets* of particulars, according to the different identity criteria adopted to conceptualize them.

At the *mereological stratum*, the IC is extensional: two entities are the same if they have the same parts. As I said, substrates belong to this stratum.

The *physical stratum* corresponds to ICs bound to spatial configuration of matter (i.e., to topo-morphological properties). Notice that these ICs are assumed to be *intensional*, in the sense that mereological identity is not always assumed as a necessary condition for physical identity. The physical stratum can be split into three separate layers:

At the *static layer*, all the non-temporal[6] properties of a particular contribute to its identity: if one of these changes, identity is lost. In this stratum only very peculiar objects are defined, namely *states* in the case of continuants and *state occurrences* in the case of occurrents. Such objects will play a crucial role in our (not yet fully developed) formal machinery, since they avoid the introduction of modal framework.

A *state* is a a sort of "frozen" continuant having a static IC, corresponding to a specific pattern of (non-temporal) properties holding for a specific amount of matter. For instance, a particular configuration of plasticine (including shape, position, color and any other non-temporal property for each single piece of plasticine) in a certain area of our table is a state.

A*world state* is a state involving a maximal amount of matter. A *world snapshot* (or simply *snapshot*) is the occurrence of a world state in a certain interval of time. A *world evolution* (or simply *world*) is a mereological sum of temporally connected snapshots. Snapshots are the ontological equivalent of temporally indexed possible worlds, and worlds are just temporal sequences of these indexed possible worlds. An occurrent *occurs*

[6] This means that we exclude properties such as "having occupied a certain location in the past", or "having been part of a certain object in the past".

in a world if it is part of it. A continuant *is defined* in a world state if its substrate exhibits the required properties for the existence of that continuant in that state. A continuant *exists* in a snapshot if it is defined in a state that occurs in that snapshot. Finally, a continuant *exists* in a world if it exists in a snapshot being part of that world.

After the static layer we have the *topological layer,* characterizing those particulars whose IC is bound to topological properties: at this layer, topological self-connection is a necessary property to maintain identity: a *piece* of matter belongs to the physical stratum (at the topological layer), while a (possibly disconnected) amount of matter belongs to the mereological stratum. The two things are *distinct* entities, since a piece of matter can cease to exist (generating new pieces) while the same amount of matter is still there.

At the *morphological layer*, the IC is bound to morphological properties (or, in general, *gestaltic* properties), like spatial shapes or temporal patterns. A change of this properties can influence identity. A cube-shapedblock is an example of an instance of this level: if its shape changes (above a certain limit) it is not *the same cube* any more, while still being the same piece of matter.

The strata above the physical stratum are related to ICs bound to properties related to the way objects interact with the external world. At the *functional stratum*, the IC is bound to functional and pragmatic properties: identity is destroyed when functionality is destroyed. At the *biological stratum*, the IC is bound to properties related to life: identity is destroyed when biological activity ceases. At the *intentional stratum*, the IC is bound to capability of intentional behavior: identity is destroyed when such capability ceases. At the *social stratum*, the IC is bound to social rules and conventions involving the interaction of intentional objects. Identity is destroyed when some of these rules change.

### 6.3 Singular and Plural Objects

Let us now examine some distinctions which only hold among concrete objects. An important one is that hodling between singular and plural objects. I shall assume that *singular objects* are those whose location is topologically self-connected (although not necessarily *maximally* self-connected). For occurrents, this means that they must extend on a continuous stretch of time, while for continuants I assume *strong* self-connection, as discussed in [Borgo *et al.* 1996]. In this way, the mereological sum of two pieces of matter touching in a point is considered as a plural object. Notice that this definition is not related with identity considerations: for instance, a person can be considered as a singular object although its identity criterion is not based (only) on topological properties.

Identity considerations play however an important role in distinguishing among different plural objects: although the simplest example of a plural object is the mereological sum of two disconnected pieces of matter, plural objects are not *just* mereological sums of disconnected singular objects, since such sums may have their own identity criteria, different from those of the objects participating to the sum (see below).

### 6.4 Bodies and Features

A further distinction among concrete objects is presented in Fig. 10. *Bodies* are objects *constituted by* a self-connected particular. *Constitution* is a primitive relation holding between a stratum and its immediate inferior, which implies dependence. For singular bodies the connection relation is the topological strong connection mentioned above, while for plural bodies *it depends on the stratum or layer where the body belongs*. For example, at the topological layer I shall adopt standard topological connection, in such a way that a lump of coal (whose single pieces are point-connected) will count as a plural body, while at the morphological layer I shall adopt some "gestaltic" notion of connection, in such a way that a constellation still counts as a plural *body*.

| | |
|---|---|
| Object | |
| Body | |
| Whole | *(a person; a lump of coal)* |
| Piece | *(a hand)* |
| Feature | *(a hole)* |

**Fig. 10**. Bodies and Features

*Features* are places marked by spatial changes of the properties of a body. I use this generic term (taken from mechanical design) to include *discontinuities* like boundaries, and *disturbances* like holes, scratches, grooves, ridges, edges, corners, spots of color, or pauses in music performances [Karmo 1977, Simons 1987, Casati and Varzi 1994]. Less obvious examples of features are certain spaces marked by intrinsically oriented objects, like the *underneath* of a table of the *front area* of a house. Differently from pieces, features are dependent on their bodies, which are called "hosts". Features belong to the more general class of *parasitic objects* like shadows [Casati and Varzi 1994], which I do not consider here.

Within bodies, we can further distinguish between *wholes* and *pieces.* A whole is a body such that: i) its constituent is *maximally* self-connected in each state where the body is defined; ii) it is independent on the existence of other bodies belonging to the same stratum. What we call "whole" here captures the important notion of Aristotelian substance, discussed for instance in [Smith 1989]. Differently from Smith, I give here a crucial importance to the quantification on all possible states: looking only at single world states, we should admit that a hand *becomes* a whole as soon as it is detached from the body, while we have good cognitive reasons to assume that the property of being a whole is rigid across all possible worlds. A hand is therefore a *piece* for us, i.e. just a body which is not a whole. A piece becomes a *fragment* when it detached from a whole.

## 7. Ontology of Universals

Let us now introduce the top-level ontology of universals, limiting ourselves to unary universals for the sake of conciseness (Fig. 11). The basic distinction is between *taxons* (which could be called also *concepts*), corresponding to sortal universals, and *properties*, corresponding to non-sortal universals. In [Guarino *et al.* 1994] we proposed a similar classification, mainly based on countability and rigidity as discriminating properties. The classification discussed below refines such a taxonomy, taking criteria of identity and individual qualities into account.

As discussed in Section 5, the difference between taxons and properties is that the former have an identity criterion associated with them: they are therefore *sortals*, in Strawson's sense. Within taxons, a preliminary distinction is made between *types* and *roles*: the former are rigid, in the sense introduced in Section 4 above and discussed in more detail in [Guarino *et al.* 1994]; the latter are not. Types are rigid classes of particulars endowed with a specific identity criterion; they can be further distinguished on the basis of the ontological properties of their instances, reflected by the top-level structure of particulars described in the previous section. We distinguish therefore between *substrate types*, i.e. classes of substrates, having an *extensional* identity criterion; *object types*, i.e. classes of objects, with an *intensional* identity criterion; and *quality types*, i.e. classes of individual qualities. According to the discussion made in [Guarino *et al.* 1994], object types are further distinguished into *substantial* and *non-substantial* types, depending on their countability. Non-substantial types have been called "categorial predicates" [Guarino *et al.* 1994], and "super-sortals" in [Pelletier and Schubert 1989].

- Unary universal
  - Taxon (Concept) (+IC)
    - Type (+R)
      - Substrate type (+Ext, -C) *(space, time, matter)*
      - Object type (-Ext)
        - Substantial — (+C) *(person)*
        - Non-substantial — (-C) *(continuant)*
      - Quality type *(color)*
    - Role (-R) *(student)*
  - Property
    - Quality attribution *(red-thing)*
    - Formal property *(self-connected)*
    - Ordinary property *(broken)*

IC = identiy criterion
Ext = extensional IC
R = rigidity
C = countability

**Fig. 11**. Top-level ontology of unary universals. The names in italics denote reified classes of particulars, appearing as instances of the metaclasses constituting the nodes of this ontology. Symbols in parentheses refer to the formal properties holding for the instances of the corresponding class.

Those unary universals which are not taxons are called (mere) *properties.* Within these, we distinguish those properties which are the result of a process of *quality attribution*, i.e. those properties which hold for an object in virtue of the fact that some of its individual quality is an instance of a certain quality type. A further class is that of *formal properties* like *atomic, independent* or *self-connected*, which characterize an individual in terms of the formal ontological properties introduced in Section 4. All other unary properties are called *ordinary properties.*

## 8. The ONTOSEEK project: using a large linguistic ontology for information retrieval

As a concrete example of the use of ontologies for information retrieval, I would like to conclude this paper reporting about OntoSeek, a general purpose tool for ontology-based information retrieval being developed by CORINTO[7] with the cooperation of LADSEB-CNR, as part of a project on retrieval and reuse of object-oriented software components [Borgo *et al.* 1997]. One of the key choices of this project has been avoiding the construction of an ontology from scratch, relying instead on a large ontology built for purposes of natural language translation. The ontology chosen, SENSUS [Knight and Luk 1994] is the result of many years of effort, and amounts to a simple taxonomic structure (no meaning axioms) of about 50,000 nodes, mostly resulting from the merging of the Wordnet thesaurus [Miller 1995] into the PENMAN top-level ontology [Bateman *et al.* 1990]. PENMAN's top-level distinctions are based on the analysis of structural invariants across multiple languages, rather than on theoretical ontological analysis like the one developed in this paper, or on "brute force" empirical effort as in the case of CYC [Guha and Lenat 1990]. The use of SENSUS ontology is an opportunity to perform an experiment of large-scale ontology reuse: its broad covering gives us the possibility to address real life applications with a moderate initial effort, making possible at the same time to verify the ontological validity of language-based distinctions on the basis of *both* concrete examples and theoretical analysis. We aim in this way to contribute to the realization of a more general, robust and well-founded top-level structure, to be used as a common *reference ontology* for multiple purposes of information interchange[8].

### 2.1 The ONTOSEEK Architecture

OntoSeek adopts a language of limited expressiveness, privileging the simplicity of use as the most important requirement. The system has been designed to:

---

[7] COnsorzio di RIcerca Nazionale Tecnologia Oggetti, as a partnership of IBM Semea, Apple Italia and Selfin SpA.

[8] Within the ANSI X3T2 committee on Information Interchange and Interpretation, and *ad-hoc group on Ontology Standard* has been recently formed, with the purpose exploring the feasibility of such a common reference ontology.

- encode an information item described by a single NL phrase into a simple graph of concepts and relations tagged by arbitrary English words;
- query a database of previously encoded information tokens by performing a process of ontology-driven *semantic match*;

We adopt a very simple graph structure for representing both queries and information data, but – differently from most of current systems – we do not assume the user to have familiarity with the vocabulary used for information encoding, relying on the ontology – supplemented by lexical information – to perform the match between queries and data.

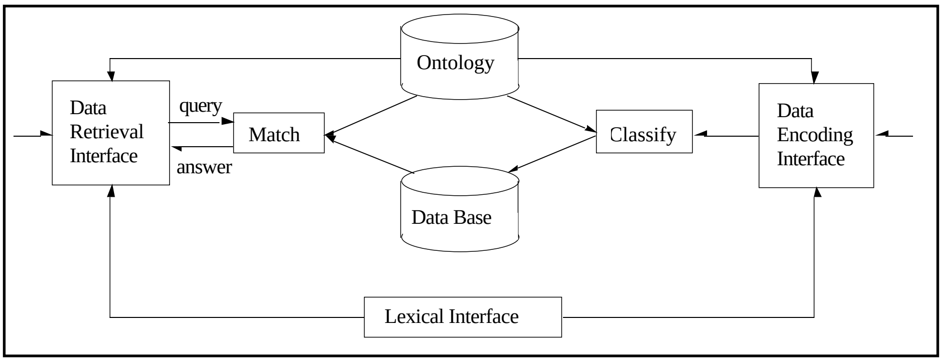


**Fig. 12**. Basic architecture of *OntoSeek* (from [Borgo *et al.* 1997]).

The basic architecture of the system is sketched in Fig. 12. In the encoding phase (which we assume to be an *interactive* process supported by NL understanding techniques), the input phrase is converted by the *Data Encoding Interface* into a simple graph like the one reported in Fig. 13, where nodes and arcs are labelled with English words. A special semantics is adopted for this graph, which is called *Lexical Conceptual Graphs* (see below). English words appearing in the graph are recognized by a *Lexical Interface* based on WordNet [Miller 1995], which asks the analyst to choose among possibly different senses associated to each word. The graph of *words* is therefore translated into a graph of *senses*, each one corresponding to a node in the *Ontology*. The sense graph is the input of a *Classification* phase, where the graph describing the component is added to the *Database*, suitably linked to the *Ontology* in order to exploit efficient search algorithms. The data retrieval process works roughly in a dual way: the query is represented again as a word graph, which is then refined into a sense graph within the *Data Retrieval Interface*. The database is then searched in order to find the information items described by those sense graphs which are subsumed by the query, according to the taxonomic constraints present in the ontology.

### 2.2 Lexical Conceptual Graphs.

Let us now briefly discuss the semantics of the graph reported in Fig. 13. The basic idea of the whole project is to exploit available lexical resources like WordNet to describe and retrieve information items. Now, most of the labels currently used in modelling formalims to denote binary relations (like "part-of", "function-of", "to-the-right-of"...) do not correspond to lexical entries. In fact, people are usually forced to invent ad-hoc relation labels when using such formalisms. Many of these labels (like "function-of", "has-part"...) are however formed from standard nouns; such nouns are indeed called "relational nouns" since they have a direct relational import, in the sense that their meaning can be fully understood only in the context of a binary relation: the noun "part" denotes the property of being a part (of something not specified), but also the range of the relation "has-part". This situation has been discussed in [Guarino 1992], where a relation like "has-part" is defined as the *relational interpretation* of the noun "part". According to that paper, only relations of this kind (i.e., relational interpretations of concepts) should be modeled as attributes (also called "roles", or "slots") *of* objects; the remaining relations should be rather modelled as constraints *between* objects.

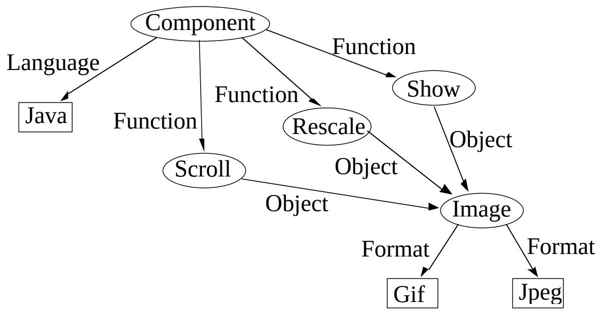


**Fig. 13**. A Lexical Conceptual Graph representing a software component (from [Borgo *et al.* 1997]).

In the light of this discussion, we can assume that arcs labelled with nouns denote the relational interpretations of such nouns, while arcs labelled with (transitive) verbs denote the relations corresponding to such verbs. In this way we are able to always guarantee a *lexical handle* to interpret the arcs appearing in our graph[9]. More formally, we call *Lexical Conceptual Graph* [10] a simplified form of conceptual graph [Sowa 1984]

---

[9] It is interesting to note that the opportunity of a linguistic restriction on role names has been advocated in a historical paper by Bill Woods [Woods 1975], who proposed a "linguistic test" which can be paraphrased as follows: A is a good role-name for X if for each filler F we can say that *F is an A of X*. See [Guarino 1992] for a full discussion.

[10] Lexical Conceptual Graphs have been called *Lexical Semantic Graphs* in [Borgo *et al.* 1997]. The present definition is more appropriate, since these graphs are simplified versions of Sowa's conceptual graphs, with a special constraint on their semantics.

like the one reported in Fig. 12, where:

1. ellipses denote concepts and rectangles denote individuals;
2. ellipses and arcs are labelled with nouns or verbs[11]; only transitive verbs are admitted for arcs;
3. if a verb appears as a label of an ellipsis, then it denotes the concept corresponding to the nominalization of that verb.
4. if a transitive verb appears as a label of an arc, it denotes the relation existing between the subject and the object of that verb.
5. if a noun C appears as a label of an arc, it denotes a relation whose range is restricted to the concept C (i.e., the relational interpretation of C). In particular, an arc with label C from the concept (or individual) A to the concept (or individual) B denotes a non-empty relation with domain A (or {A}) and range C ∩ B (or C ∩ {B}) (Fig. 14).

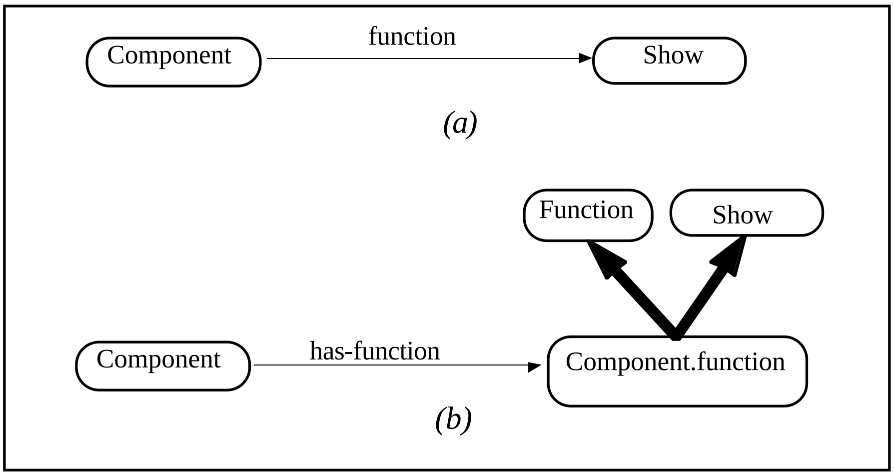


**Fig. 14.** Semantics of Lexical Conceptual Graphs (LCG). The LCG in (a) is a shorthand for the ordinary semantic graph reported in (b). The advantage is that the relation "has-function" is represented by means of the lexical item "function" (adapted from [Borgo *et al.* 1997]).

### 2.3 The advantages of lexical-driven information retrieval

Let us now discuss the approach presented above in the light of the current literature, with special regard to systems based on faceted classification schemes. With respect to these approaches, our main points are:

[11] Infinitive forms are supposed to be abbreviations for the corresponding nominalizations: "show" denotes the concept of "showing" (whose instances are events).

- a simple (but semantically rigorous) representation language of intermediate expressivity (roughly situated between facets and description logics);
- a very large choice of description terms both in the encoding and retrieval phase, due to the use of full English lexicon combined with a large ontology;
- the possibility of semantic checks guided by the ontology;

Regarding facets-based schemes, a well known limitation is related to the fixed number of descriptors (keywords). When we consider a large repository of non-homogeneous data, the choice of such descriptors can be very difficult. Classifying these data in a satisfactory way implies defining a big set of descriptors, whose exact meaning is difficult to control, document and maintain. It may even happen that the intended meaning of a descriptor varies depending on data types. When dealing with homogeneous data, the importance of this problem decreases, but it does not disappear. Suppose we have a repository of products of similar kind: a few facets may be enough to describe it, but the keywords to be allowed under each facets may be quite many.

In the lexicon-driven approach we have described these problems are overcome. The classification is flexible enough to describe large non-homogeneous repositories, where it is necessary to represent a variety of coarse distinctions, as well as homogeneous repositories, where fine-grained distinctions are important. Moreover, the ontology (supported by the lexical interface) makes possible to clarify the meaning of each term used.

In sum, the real importance of facets-based classification systems lies in the particular structure of keywords they propose, designed to be effective for the information retrieval task. Within our approach we can still use this structure as a *template*, offering however to the user three extra-services: the possibility to *paraphrase* the description, the possibility to *add meaning* to the description, and the possibility to *better understand* the meaning of the words adopted. In this perspective, a lexicon-driven system can be seen as a smooth but powerful improvement with respect to a facets-based system.

Of course, the system we have discussed has many limitations, which need to be taken into due account. As an example taken from our application in the domain of software components, consider an item called “RAM doubler” which does not magically double the physical RAM (unfortunately), but does enrich the performance of a computer. If the software analyst uses the very same words, “Ram” and “double”, to classify this software, then a query like “enhancing performance”, would never match the above object. Note that even in the case where the item would be described by the expression “enhancing memory”, the query could not be satisfied, unless knowing that the performance of a computer can be enhanced by increasing the perfomance of a *part* of it, namely its memory. This kind of problem is faced in each approach based on just natural language descriptions, where no background domain knowledge is available.

## Acknowledgements

I hope to have offered a contribution showing the possible advantages of philosophical distinctions in the practice of knowledge engineering, and information retrieval and extraction in particular. Many of these distinctions still present serious philosophical problems, but I believe they can be however very useful for practical applications. Hopefully, their introduction in engineering systems can stimulate further fundamental research.

This work has been partially funded by a contract between LADSEB-CNR and the Apple/IBM Consortium “CORINTO”. I am grateful to the co-authors of the cited papers (Stefano Borgo, Massimiliano Carrara, Pierdaniele Giaretta, Claudio Masolo, Guido Vetere) for their support.

## Bibliography

Albertazzi, L. 1989. *Strati.* Riverdito, Trento.

Artale, A., Franconi, E., Guarino, N., and Pazzi, L. 1996. Part-Whole Relations in Object-Centered Systems: an Overview. *Data and Knowledge Engineering*, **20**(3): 347-383.

Aune, B. 1991. Metaphysics of Analytic Philosophy. In H. Burkhardt and B. Smith (eds.), *Handbook of Metaphysics and Ontology*. Philosophia, Munich: 539-543.

Bateman, J. A., Kasper, R. T., Moore, J. D., and Whitney, R. A. 1990. A General Organization of Knowledge for Natural Language Processing: the PENMAN upper model. USC/Information Sciences Institute, Marina del Rey, CA.

Borgo, S. and Guarino, N. 1997. An Ontological Theory of Physical Objects. In *Proceedings of 1997 Workshop on Qualitative Reasoning (QR 97)*. Cortona, Italy.

Borgo, S., Guarino, N., and Masolo, C. 1996. A Pointless Theory of Space based on Strong Connection and Congruence. In *Proceedings of Principles of Knowledge Representation and Reasoning (KR96)*. Boston, MA, Morgan Kaufmann: 220-229.

Borgo, S., Guarino, N., and Masolo, C. 1996. Stratified Ontologies: the Case of Physical Objects. In *Proceedings of ECAI-96 Workshop on Ontological Engineering*. Budapest.

Borgo, S., Guarino, N., Masolo, C., and Vetere, G. 1997. Using a Large Linguistic Ontology for Internet-Based Retrieval of Object-Oriented Components. In *Proceedings of 1997 Conference on Software Engineering and Knowledge Engineering*. Madrid, Knowledge Systems Institute, Snokie, IL, USA.

Bouaud, J., Bachimont, B., Charlet, J., and Zweigenbaum, P. 1995. Methodological Principles for Structuring an "Ontology". In *Proceedings of IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing*. Montreal, Quebec, Canada.

Burkhardt, H. and Smith, B. (eds.) 1991. *Handbook of Metaphysics and Ontology*. Philosophia Verlag, Münich.

Casati, R. and Varzi, A. (eds.) 1996. *Events*. Dartmouth, Aldershots, USA.

Casati, R. and Varzi, A. C. 1994. *Holes and Other Superficialities.* MIT Press/Bradford

Books; revised paperback edition 1995., Cambridge (MA) and London (UK). Ed. Italiana: Buchi ed altre superficialita' (Trad. it. di Libero Sosio), Milano, Garzanti 1996.

Cleland, C. 1991. On the Individuation of Events. *Synthese*, **86**: 229-254. Reprinted in R. Casati and A. Varzi (eds.), *Events*, Dartmouth 1996.

Cocchiarella, N. B. 1991. Formal Ontology. In H. Burkhardt and B. Smith (eds.), *Handbook of Metaphysics and Ontology*. Philosophia Verlag, Munich: 640-647.

Cowie, J. and Lehnert, W. 1996. Information Extraction. *Communications of the ACM*, **39**(1): 80-91.

Genesereth, M. R. and Nilsson, N. J. 1987. *Logical Foundation of Artificial Intelligence*. Morgan Kaufmann, Los Altos, California.

Griffin, N. 1977. *Relative Identity*. Oxford University Press, Oxford.

Gruber, T. 1995. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human and Computer Studies*, **43**(5/6): 907-928.

Gruber, T. R. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, **5**: 199-220.

Guarino, N. 1992. Concepts, Attributes and Arbitrary Relations: Some Linguistic and Ontological Criteria for Structuring Knowledge Bases. *Data & Knowledge Engineering*, **8**: 249-261.

Guarino, N. 1994. The Ontological Level. In R. Casati, B. Smith and G. White (eds.), *Philosophy and the Cognitive Science*. Hölder-Pichler-Tempsky, Vienna: 443-456.

Guarino, N. 1995. Formal Ontology, Conceptual Analysis and Knowledge Representation. *International Journal of Human and Computer Studies*, **43**(5/6): 625-640.

Guarino, N. 1996. Understanding, Building and Using Ontologies. LADSEB Internal Report 6/96.

Guarino, N. 1997a. Some Organizing Principles for a Unified Top-Level Ontology. In *Proceedings of AAAI Spring Symposium on Ontological Engineering*. Stanford, CA, AAAI Press.

Guarino, N. 1997b. Understanding, Building, and Using Ontologies: A Commentary to "Using Explicit Ontologies in KBS Development", by van Heijst, Schreiber, and Wielinga. *International Journal of Human and Computer Studies (in press)*(46): (in press).

Guarino, N., Carrara, M., and Giaretta, P. 1994. Formalizing Ontological Commitment. In *Proceedings of National Conference on Artificial Intelligence (AAAI-94)*. Seattle, Morgan Kaufmann: 560-567.

Guarino, N., Carrara, M., and Giaretta, P. 1994. An Ontology of Meta-Level Categories. In D. J., E. Sandewall and P. Torasso (eds.), *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourth International Conference (KR94)*. Morgan Kaufmann, San Mateo, CA: 270-280.

Guarino, N. and Giaretta, P. 1995. Ontologies and Knowledge Bases: Towards a Terminological Clarification. In N. Mars (ed.) *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing 1995*. IOS Press, Amsterdam: 25-32.

Guarino, N. and Poli, R. (eds.) 1995. *Formal Ontology in Conceptual Analysis and Knowledge Representation*. Special issue of the International Journal of Human and Computer Studies, vol. 43 n. 5/6, Academic Press.

Guha, R. V. and Lenat, D. B. 1990. CYC: a Mid-Term Report. *AI Magazine*, **11**(3): 32-59.

Hirsch, E. 1982. *The Concept of Identity*. Oxford University Press, New York, Oxford.

Hirst, G. 1991. Existence Assumptions in Knowledge Representation. *Artificial Intelligence*, **49**: 199-242.

Karmo, T. 1977. Disturbances. *Analysis*, **37**: 147-148.

Knight, K. and Luk, S. 1994. Building a Large Knowledge Base for Machine Translation. In *Proceedings of American Association of Artificial Intelligence Conference (AAAI-94)*. Seattle, WA.

Lombard, L. B. 1979. Events. *Canadian Journal of Philosophy*, **9**(3). Reprinted in Roberto Casati and Achille Varzi (eds.), Events, Dartmouth 1996.

Mahesh, K. 1996. Ontology Development for Machine Translation: Ideology and Methodology. New Mexico State University, Computing Research Laboratory MCCS-96-292.

Miller, G. A. 1995. WORDNET: A Lexical Database for English. *Communications of ACM*(11): 39-41.

Mulligan, K. (ed.) 1992. *Language, Truth and Ontology*. Kluwer, Dordrecht.

Mulligan, K., Simons, P. M., and Smith, B. 1984. Truth Makers. *Philosophy and Phenomenological Research*, **44**: 287-321.

Noonan, H. (ed.) 1993. *Identity*. Dartmouth, Aldershot, USA.

Pelletier, F. J. and Schubert, L. K. 1989. Mass Expressions. In D. Gabbay and F. Günthner (eds.), *Handbook of Philosophical Logic*. Reidel: 327-405.

Poli, R. 1995. Bimodality of Formal Ontology and Mereology. *International Journal of Human and Machines Studies*, **43**(5/6): 687-696.

Poli, R. 1996. Ontology and Knowledge Organization. In *Proceedings of 4th Conference of the International Society of Knowledge Organization (ISKO 96)*. Washington.

Simons, P. 1987. *Parts: a Study in Ontology*. Clarendon Press, Oxford.

Smith, B. (ed.) 1982. *Parts and Moments: Studies in Logic and Formal Ontology*. Philosophia Verlag, München.

Smith, B. 1989. The Primacy of Place: an Investigation in Brentanian Ontology. *Topoi*, **8**: 43-51.

Smith, B. and Mulligan, K. 1983. Framework for Formal Ontology. *Topoi*, **2**: 73-85.

Sowa, J. 1995. Top-level ontological categories. *International Journal of Human and Computer Studies*, **43**: 669-685.

Strawson, P. F. 1959. *Individuals. An Essay in Descriptive Metaphysics*. Routledge, London and New York.

Swartout, B., Patil, R., Knight, K., and Russ, T. 1996. Toward distributed use of large-scale ontologies. In *Proceedings of 10th Knowledge Acquisition for Knowledge-Based Systems Worskhop*. Banff, Canada.

Uschold, M. and Gruninger, M. 1996. Ontologies: Principles, Methods and Applications. *The Knowledge Engineering Review*, **11**(2): 93-136.

Van Heijst, G., Schreiber, A. T., and Wielinga, B. J. 1997. Roles aren't Classes: a Reply to Nicola Guarino. *International Journal of Human and Computer Studies*(in press).

Van Heijst, G., Schreiber, A. T., and Wielinga, B. J. 1997. Using Explicit Ontologies in KBS Development. *International Journal of Human and Computer Studies*(in press).

Varzi, A. 1996. Parts, Wholes, and Part-Whole Relations: The Prospects of Mereotopology. *Data and Knowledge Engineering*, **20**(3): 259-286.

Winston, M., Chaffin, R., and Herrmann, D. 1987. A Taxonomy of Part-Whole Relations. *Cognitive Science*, **11**: 417-444.

Woods, W. 1986. Important issues in knowledge representation. *Proceedings of the IEEE*, **74**(10): 1322-1334.

Woods, W. A. 1975. What's in a Link: Foundations for Semantic Networks. In D. G. Bobrow and A. M. Collins (eds.), *Representation and Understanding: Studies in Cognitive Science*. Academic Press: 35-82.

*Summer School on Information Extraction*, Frascati, July 14-19, Frascati, Italy.
To appear in a volume edited by M. T. Pazienza, published by Springer Verlag.

## Appendix 1. A *preliminary* top-level for the ontology of particulars

NOTE - This ontology is in no sense stable and complete. Formal axioms are intended to characterize the meaning of the concepts used, but they are not reported here due to the preliminary nature of this work.

### Notation

1. Each line denotes a node in the taxonomy. Parenthesized expressions in regular style denote alternative names for nodes.
2. Indentation represents subsumption.
3. The symbol — is used to avoid the repetition of the term used at tbe above level.
4. Lines in italics always denote instances of the concept denoted by the above level.
5. The direct children of a node denote mutually exclusive classes. Such classes are also intended as exhaustive if the node "..." does not appear as a child.

## Particulars

- **Substrate**
  - Location
    - Space
    - Time
  - Matter
    - Iron
    - Wood
    - …
- **Object**
  - Concrete object
    - Continuant
      - Singular —
        - Singular body
          - Physical body
            - Topological body
              - Topological whole
                - *(a planet)*
              - Topological piece
                - *(a piece of wood)*
            - Morphological body
              - Morphological whole
                - *(a cube of wood)*
              - Morphological piece
                - *(a mountain)*
            - Functional body
              - Functional whole
                - Artifact

…
Functional piece
Artifact component
Organic structure
…
Singular feature
Physical feature
Topological feature
*(the boundary of a blob)*
Morphological feature
*(the top of a mountain)*
*(a hole in a piece of cheese)*
Functional feature
*(the head of a bolt)*
*(the profile of a wing)*
Plural —
Plurality (Plural body)
Physical plurality
Mereological plurality
Topological plurality
*(a lump of coal)*
Morphological plurality
*(a line of trees)*
*(a pattern of stars)*
Functional plurality
*(a disassembled artifact)*
Plural feature
*(a hole in a lump of coal)*
*(the sword of Orion)*
Occurrent
Static occurrent
World state
…
Topological occurrent
Morphological occurrent
Functional occurrent
Biological occurrent
Intentional occurrent
Action
Mental event
…
Social occurrent
Communicative event
…
Abstract object

**Quality**

*(the color of a rose)*
*(the style of a talk)*